HECFSIO Panel

Metadata

Data Volume is Increasing

- But what about metadata … ?

Data Volume is Increasing

- We'd love it if as the data grows, the metadata stays the same

OrangeFS

Data Volume is Increasing

- We would be happy if the metadata grew at the same rate as the data

OrangeFS

Data Volume is Increasing

- What we are seeing is that the metadata is growing at a faster rate than the data

OrangeFS

Metadata Growth

- Not just bigger files
  - More files
  - Smaller files
  - Files that characterize and/or index other files
- More of it, and also more complex
- The metadata IS the data
- File systems need more
  - Replication, HSM, caching, etc

The File System As Index

- Many applications build large directory trees with many files in each directories specifically to locate data.
  - Significant amounts of traditional metadata is wasted – in many cases owner, group, access permissions are the same for a large number of entries.
  - Opportunities or organize storage may be lost
  - WHY don't they use a DBMS?

What Can the File System Do?

- First, ask ourselves *should* we do anything?
  - Some would argue this is outside of our purview
- In many cases we already have solutions
  - Indexed directories
    - OFS uses BerkeleyDB b-trees
  - Distributed directories (eg. Giga+)
    - Extensible hashing speeds lookup
    - Parallel access to the directories
  - On-server indexing could be applied to user data

"Light" Directories

- Allow a "striped-down i-node"
  - Files in light directory actually part of larger data set
  - Store only what metadata is critical to each file
- Alternatively …
  - Provide a user interface for building a tree-structured index using the existing file system commands
  - Data in one large file with multiple entry points via the index

Indexing User data

- OrangeFS prototype allows extended attributes on files to be searched using Berkeley DB
- Allow users to build an on-server index to their own files

The Key Is Interfaces

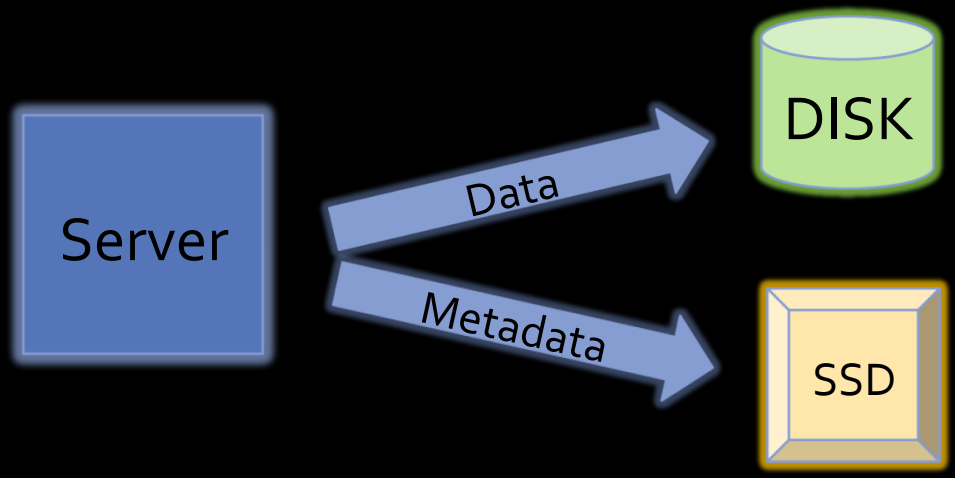
- Why don't users use the existing DB infrastructure?
- So many researchers are turning to computing – many don't know DBMS exist, most don't know how to use it
- Can we develop interfaces that
  - They will use
  - Don't wreck our standards
  - Give us room to grow further

A Revolutionary Approach

- Programming model with a Global Address Space like UPC, ParalleX, …
- Unify the persistent storage name space with the GAS
- Metadata becomes attributes of data objects created by user, stored, indexed, and used for retrieval
- PXFS – PVFS for ParalleX


OrangeFS

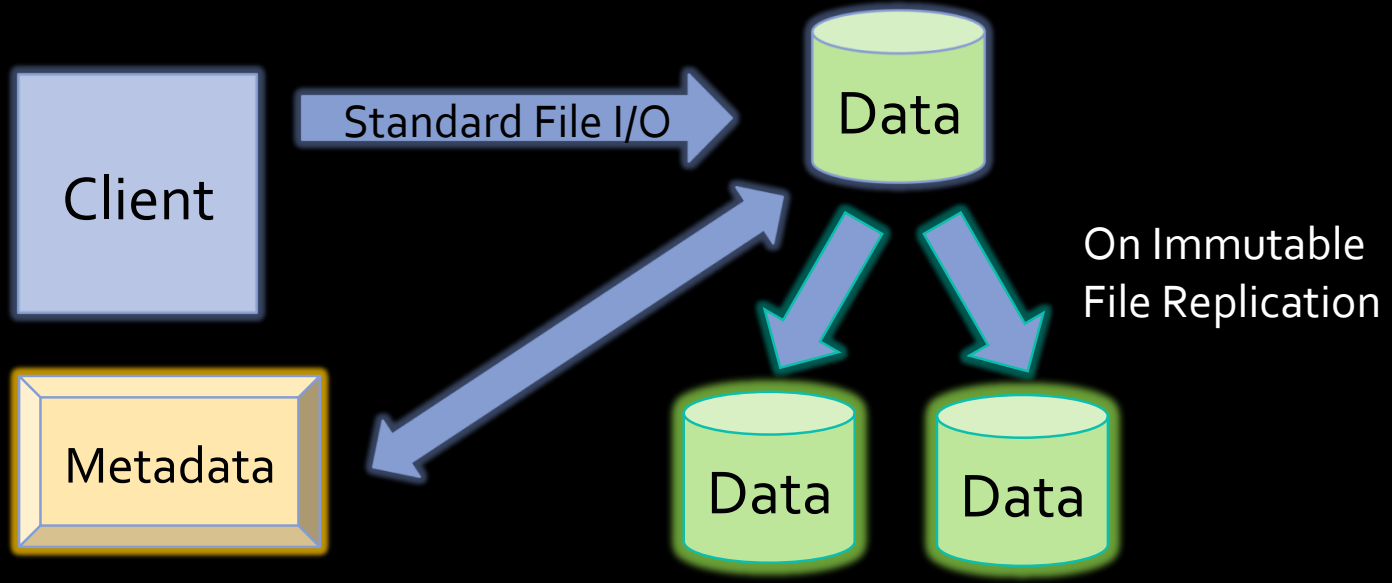
SSD Metadata Storage



- Frequent metadata updates to storage
  - Improves Reliability
  - Reduces Performance
- Writing metadata to SSD
  - Improves Performance
  - Maintains Reliability

OrangeFS

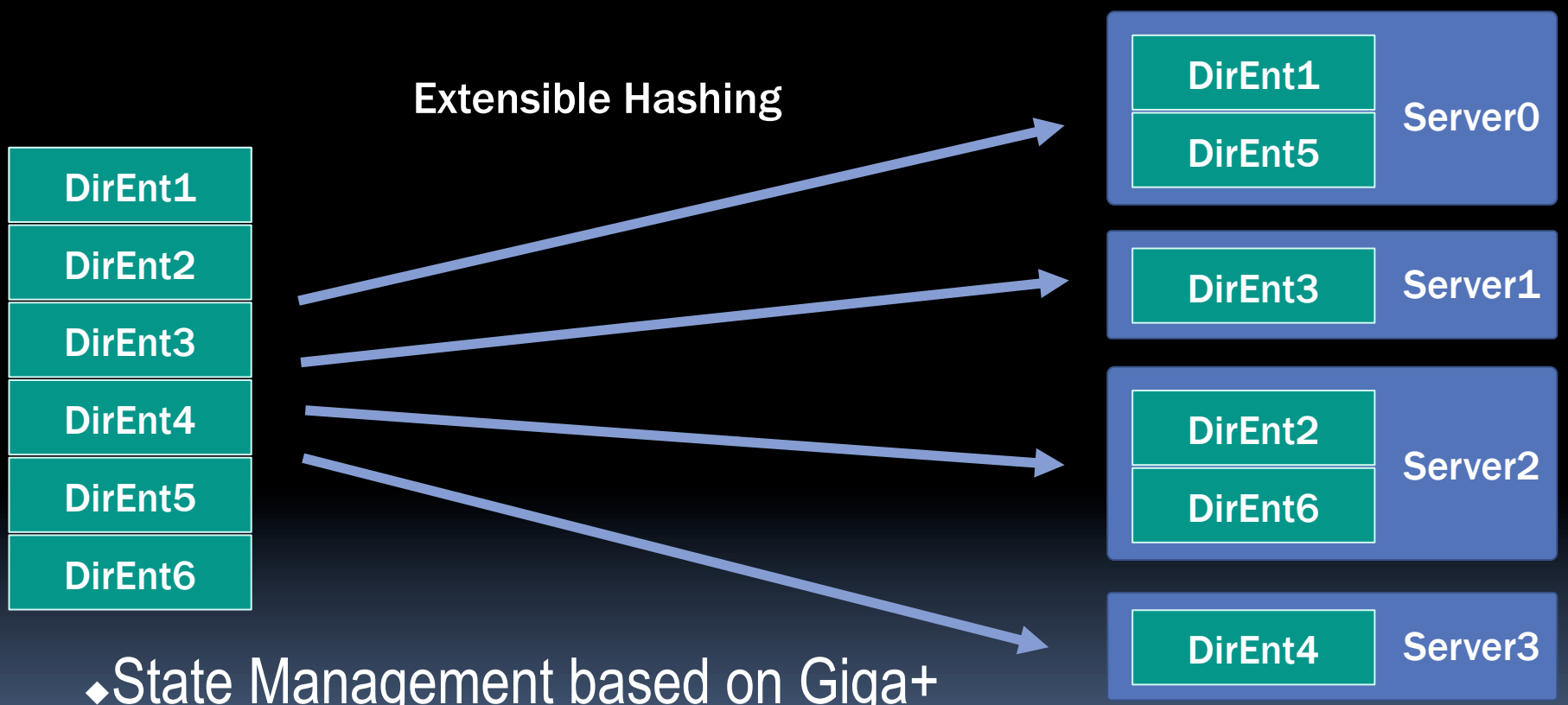
Replicate On Immutable



- Replicate data to provide resiliency
  - Initially replicate on Immutable
  - Client read fails over to replicated file if primary is unavailable

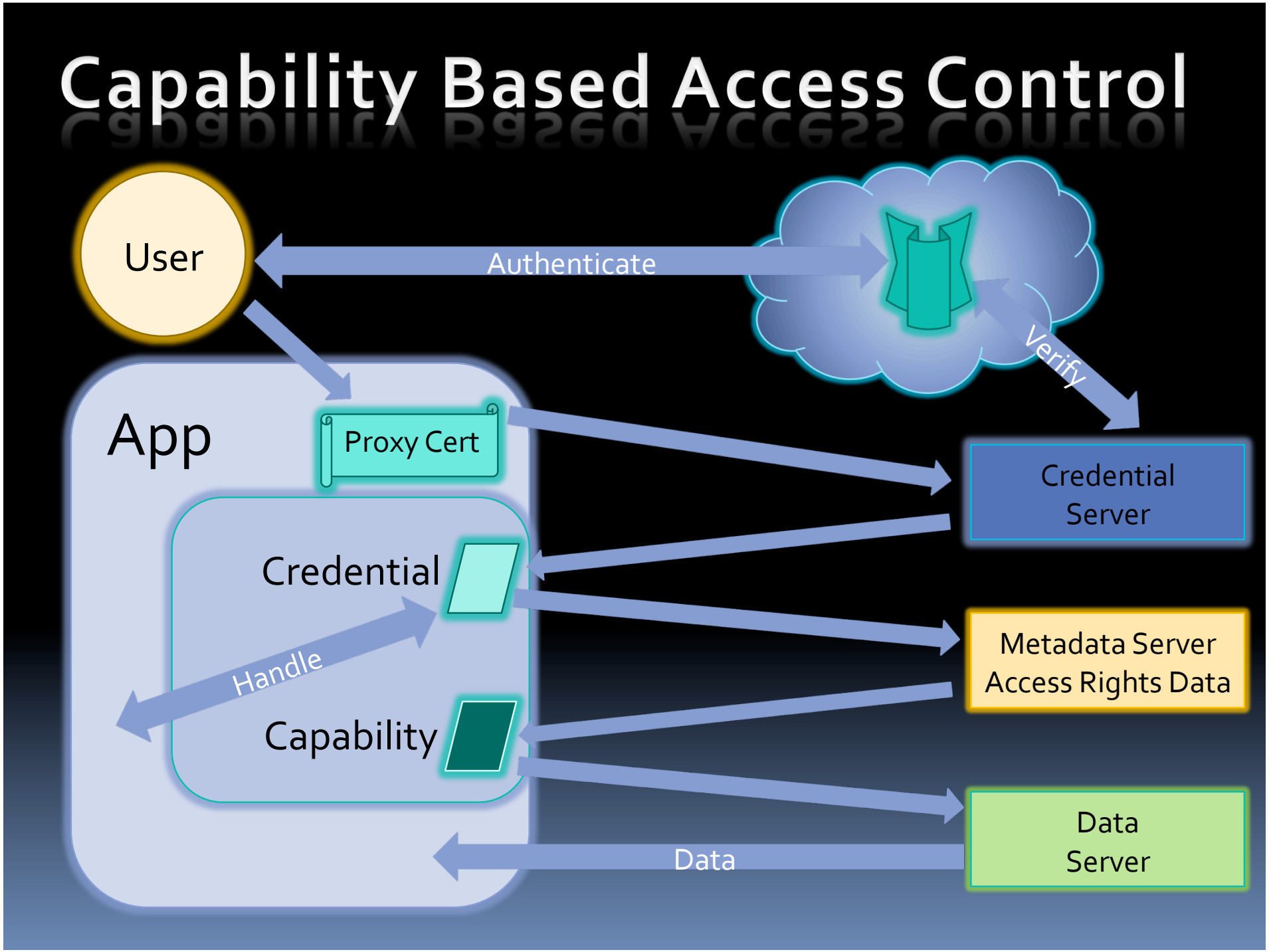
OrangeFS

Distributed Directories

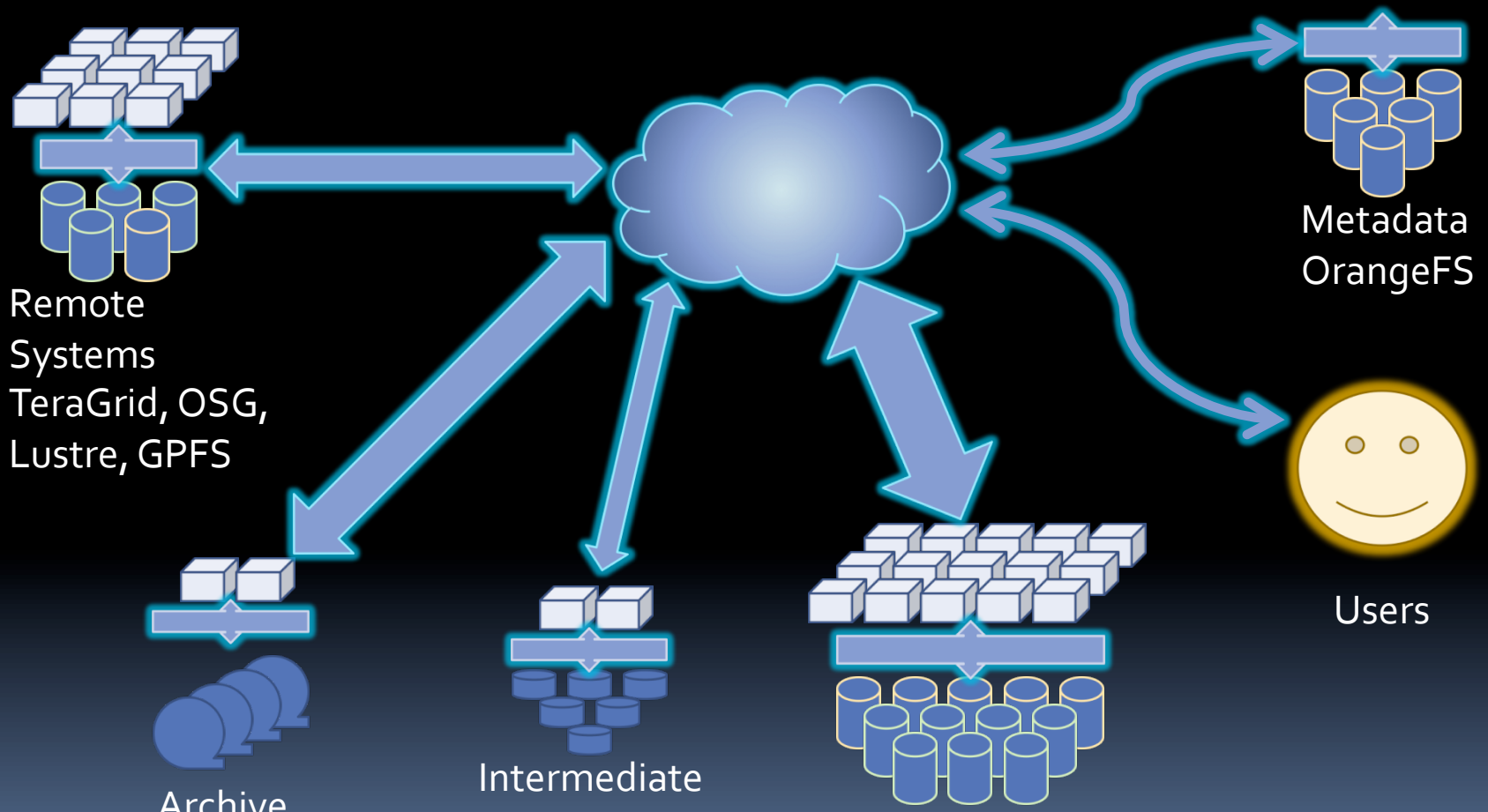


- State Management based on Giga+
- Improves access times for directories with a very large number of entries

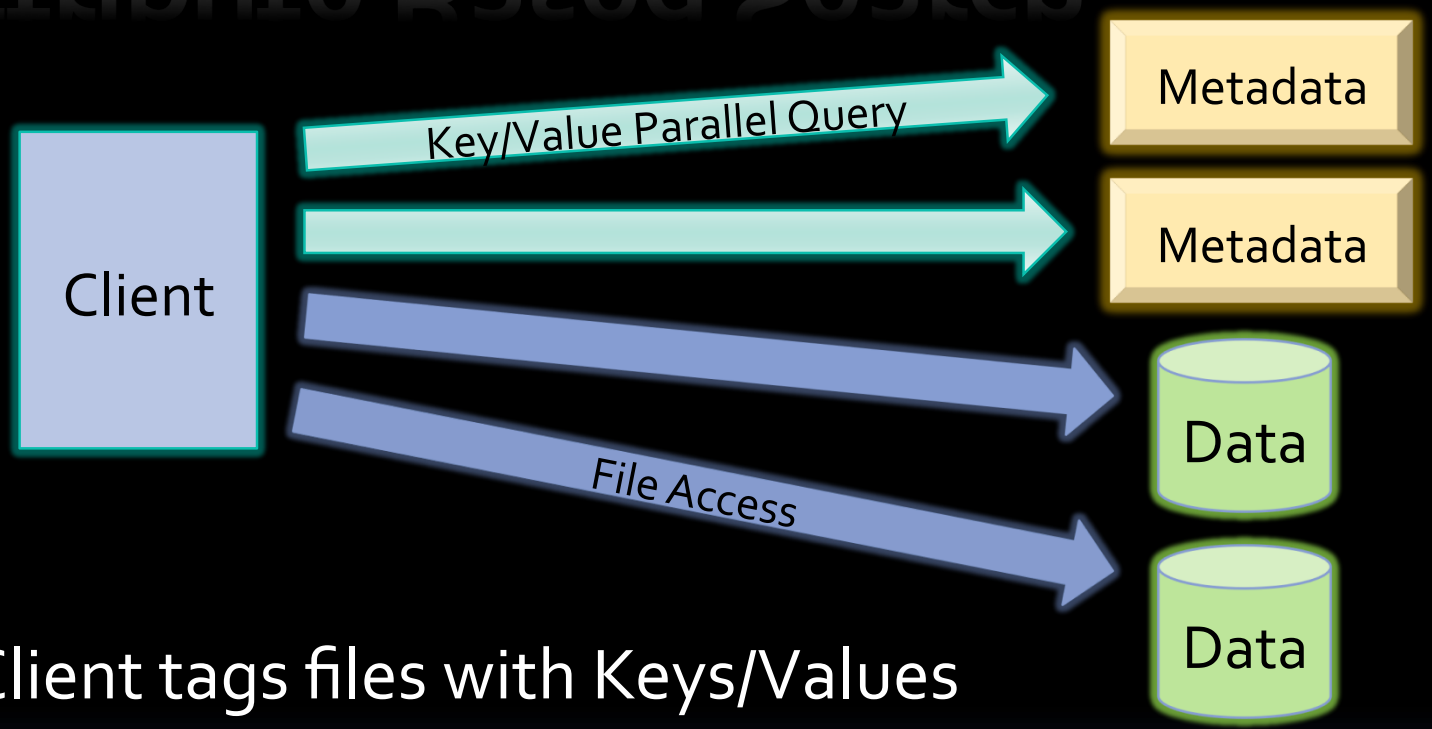
OrangeFS


Capability Based Access Control
User
Authenticate
Verify
App
Proxy Cert
Credential
Server
Credential
Handle
Metadata Server
Access Rights Data
Capability
Data
Server
Data

Hierarchical Data Management



Attribute Based Search



- Client tags files with Keys/Values
- Keys/Values indexed on Metadata Servers
- Clients query for files based on Keys/Values
- Returns file handles with options for filename and path

OrangeFS